Detección de pulsaciones por minuto (BPM) en tiempo real a partir de un Sistema de Inteligencia Artificial para el género musical Salsa

Murcia Gomez Danny Julian,
julianmurcia1@gmail.com
Pontificia Universidad Javeriana de Cali
Santiago de Cali – Valle del Cauca
Calle 18 # 118 - 250

Abstract The objective of this document is about to show the use of artificial intelligence to predict the beats per minute (BPM) only for the salsa music genre. Starting from a photo of a melspectrogram with a frequency range between 400 Hz and 2048 Hz as input to the neural network, we obtain the beats per minute of the songs as a prediction of the neuronal network. A total of 3 different neuronal networks are trained and subsequently we analyze the results individually. An excellent result was obtained in one of the neuronal networks, obtaining an accuracy of 85% in the test data.

***Palabras Clave*— Machine Learning, Música, Inteligencia Artificial, Detección de Pulsaciones por minuto.**

I. Introducción

En la actualidad plataformas como Spotify, iMusic, Deezer y otras registran en promedio más de 30 millones de canciones [1] y más de 100 millones de usuarios usando este servicizo [2]. Al haber tal cantidad de datos en canciones, se busca hallar modelos que faciliten la interpretación de ondas de sonido a datos, para facilitar nuevos controles sobre la música, como son los modelos de estilo, análisis de armonía, similitud tímbrica, detección del beat, entre otros. El género musical salsa, es uno de los géneros musicales que presenta poco análisis en el campo científico por su gran variedad de ritmos caribeños que lo hacen uno de los géneros con mayor cantidad de combinaciones de géneros musicales.

El BPM (Beat per Minute) es una unidad empleada para medir el ritmo en música; usado por científicos, productores y mezcladores de música que lo aprovechan para el análisis, reproducción o mezcla de la música [3]. Pero al género musical salsa tiene unos ritmos complejos y variados, causando en los modelos de detección de pulso actuales incertidumbre, siendo incapaces de detectar en tiempo real a que BPM (Beat per Minute) va el género musical Salsa.

La finalidad del proyecto es encontrar en el género musical salsa patrones en las canciones para entrenar una red neuronal que sea capaz de clasificar el BPM de una canción entre un rango de 60 BPM hasta 220 BPM y crear una vista amigable para un usuario final que pueda hacer uso de la Inteligencia artificial de una manera rápida, fácil y sencilla.

II. Fundamentación teórica

A. Conjunto de datos

El conjunto de datos se recolectó de manera manual, de una base de datos de colección privada de 1119 canciones de salsa y de la base de datos Extended-

BIGROOMDATASET [4] que contiene 47 canciones de salsa, para un total de 185730 minutos de canciones únicamente de Salsa. La distribución de los datos fueron los siguiente:

- Conjunto de datos de entreno.
- Conjunto de datos de validación.
- Conjunto de datos de prueba.
- Conjunto de datos de Extended-BIGROOM.

B. *Explicación del uso del Rango de Hz*

El cencerro es una pequeña campana que es hecha a base de cobre o acero; su tono varía dependiendo del tamaño de la campana y donde se toque. Se caracteriza por ser un idiófono (Su tono es único), por lo tanto, su sonido siempre va a ser indeterminado (no es constante en un rango de Hz [5]). En determinados momentos este instrumento en la salsa puede llevar el ritmo y velocidad; y al ser un instrumento con sonido único, es fácilmente diferenciable de los demás.

Se realizó pequeños experimentos donde se grabó el sonido de un cencerro tocando el instrumento en diferentes partes, aquí los resultados:

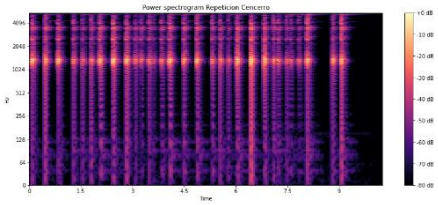

Figura 2.8: Imagen de un Power Spectrogram de un Cencerro tocado repetidamente. Autor: Freddythehat de Wikipedia. Fecha: 19 de febrero del 2007. Propósito únicamente investigativo.

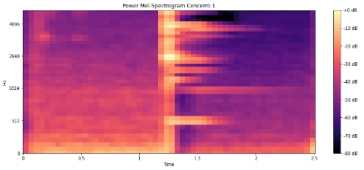

Figura 2.9: Imagen de un Power Mel-Spectrogram de un Cencerro tocado únicamente en la parte superior.

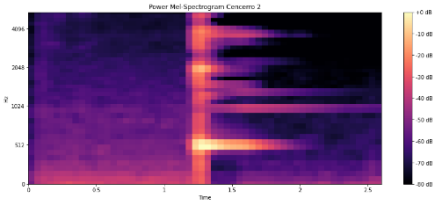

Figura 2.10: Imagen de un Power Mel-Spectrogram de un Cencerro tocado únicamente en la parte del medio.

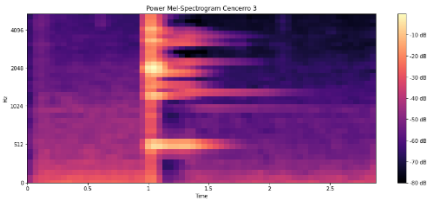

Figura 2.11: Imagen de un Power Mel-Spectrogram de un Cencerro tocado únicamente en la parte baja.

Concluimos de las imágenes 2.8, 2.9, 2.10 y 2.11 que el cencerro es un instrumento que más o menos entre 400 Hz hasta los 2048 Hz se pueden encontrar al menos una activación. Si se escoge un rango mayor a 2048 Hz se vuelve muy complicado captar la activación, gracias a que en este género los diferentes instrumentos y voces en la canción pueden interferir.

C. *Redes Neuronales*

1) *Entradas*

Las entradas consisten en fotos que se caracterizan por tener 1 canal de color con 256 columnas con 40 filas. Son 256 columnas que hacen referencia al tiempo de la canción, dónde 256 columnas de tiempo es equivalente a un aproximado de 11.85 segundos. Son 40 filas que hacen referencia a las unidades de frecuencia. Tiene un canal de color, el cual está en un rango de 0 a 1. Las entradas hacen referencia a un melspectrogram [6] (imagen 2.7).

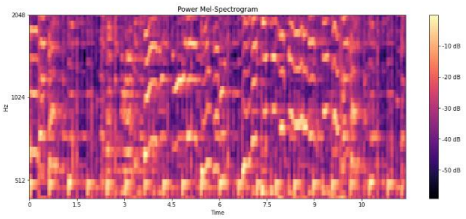

Figura 2.7: Imagen de un Power Mel-Spectrogram de un pedazo de la canción Horas Lindas de Adolecentes Orquesta. Entrada de la red neuronal.

2) Salidas

Tenemos un total de 160 salidas que nos representan los BPM desde 60 BPM siendo la posición0 en el arreglo de salida hasta 220 BPM que representa la posición 160 del arreglo de salida. Las redes neuronales que se crearon nos predicen una salida en 11.85 segundos de un Tempo local.

3) Modelo I

La red neuronal la podemos observar en la imagen 2.12. Este modelo cuenta con 5’094,578 parámetros entrenables.

Esta red neuronal se reduce de la siguiente red neuronal [7].

4) Modelo II

La segunda red neuronal la podemos observar en la imagen 2.14. Este modelo cuenta con 1’080,752 parámetros entrenables.

Esta red neuronal es sencilla y simple; se basa en el modelo de la siguiente red neuronal [8]

5) Modelo III

La tercera red neuronal la podemos observar en la imagen 2.16. Este modelo cuenta con 8’186,656 parámetros entrenables.

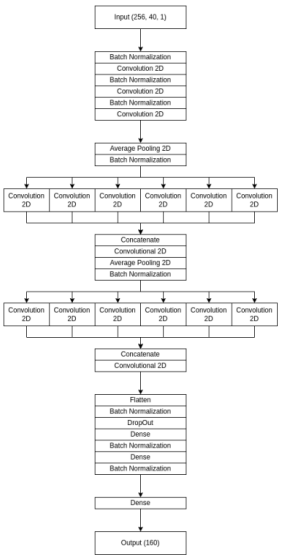

Figura 2.12: Imagen del modelo de la red neuronal 1

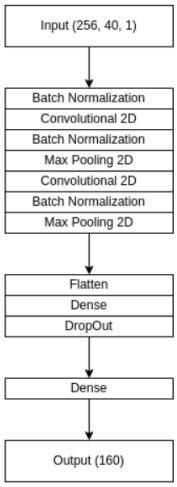

Figura 2.14: Imagen del modelo de la red neuronal 2.

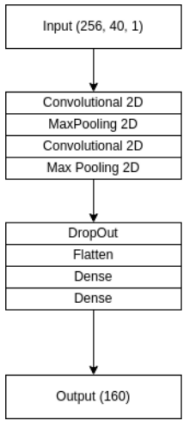

Figura 2.16: Imagen del modelo de la red neuronal 3

D. Entrenamiento

Se usó aproximadamente el 90 % de los datos para entrenar en la red y aproximadamente un 10 % para validación; los datos son randomizados. Se usa para su función de precisión “Sparse Categorical crossentropy” [9] que nos permite obtener un resultado con muchos tipos de clasificación y una función de perdida Adam [10].

Cada modelo tiene sus propias características para entrenar: el Modelo1 se entrenó por 20 generaciones y 20 imágenes por ciclo, el Modelo 2 se entrenó con 30 generaciones y 20 imágenes por ciclo, y finalmente el Modelo 3 se entrenó 20 generaciones y 10 imágenes por ciclo.

III. Resultados

A. Modelo I

Los resultados fueron los siguientes. Datos de entrenamiento: Obtuvo una perdida aproximada de 0.6535 y una predicción del 88.49%. Datos de Validación: Obtuvo una perdida aproximada de 2.3211 y una predicción del 51.57%. Datos de Prueba: Obtuvo una perdida aproximada de 1.0039 y una predicción del 84.93%. Datos de Extended-BIGROOM: Obtuvo una perdida aproximada de 8.4479 y una predicción del 9.57%.

B. Modelo II

Los resultados fueron los siguientes. Datos de entrenamiento: Obtuvo una perdida aproximada de 0.0222 y una predicción del 99.63%. Datos de Validación: Obtuvo una perdida aproximada de 0.5619 y una predicción del 92.83%. Datos de Prueba: Obtuvo una perdida aproximada de 6.5750 y una predicción del 12.87%. Datos de Extended-BIGROOM: Obtuvo una perdida aproximada de 9.3137 y una predicción del 1.06%.

C. Modelo III

Los resultados fueron los siguientes. Datos de entrenamiento: Obtuvo una perdida aproximada de 0.0529 y una predicción del 99.24%. Datos de Validación: Obtuvo una perdida aproximada de 0.6198 y una predicción del 92.60%. Datos de Prueba: Obtuvo una perdida aproximada de 8.6703 y una predicción del 7.85%. Datos de Extended-BIGROOM: Obtuvo una perdida aproximada de 22.0336 y una predicción del 2.13%.

Concluimos con que el desempeño para el conjunto de datos de Extended-BIGROOM es extremadamente pobre, podemos asumir que no encontró ningún patrón dentro de los datos.

D. Deteccion de BPM

Los resultados obtenidos por las redes neuronales no detectan un BPM global de la canción, si no que detectan un tempo local de la canción, por lo que es necesario pasar varias veces la canción en diferentes puntos de reproducción para así obtener un estimado de BPM global.

La canción a la cual se quiera detectar el BPM se parte en 6 partes diferentes y se hace un promedio de sus resultados; los resultados al ser un arreglo de 160 datos donde cada posición se refiere a un BPM (si está en la posición 0 se refiere a que tiene un BPM de 60) después de promediar los resultados se obtiene un BPM global.

IV. Interfaz Grafica de Usuario

A. Encuesta

Se creó una encuesta con el propósito de que las personas introduzcan manualmente el BPM, indicando donde este cada tempo de la canción. Se puede observar en la imagen 2.20.

Figura 2.20: Imagen de la encuesta final.

Este desarrollo tiene el siguiente diagrama de flujo para el usuario:

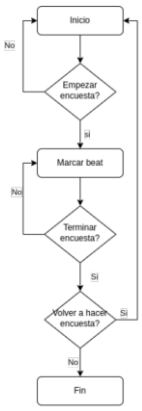

Figura 2.25: Diagrama de flujo de la vista Encuesta.

B. Predicción de BPM

Se crea una interfaz de usuario amigable y sencilla donde el usuario selecciona la canción que quiere hallar el BPM (Imagen: 2.27).

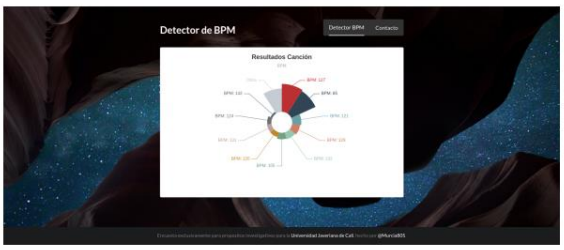

Figura 2.27: Imagen de la interfaz de Prediccion resultado final.

Este desarrollo tiene el siguiente diagrama de flujo para el usuario:

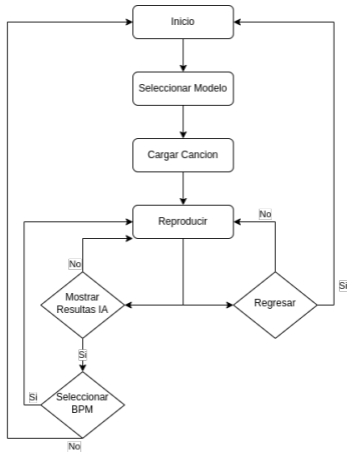

Figura 2.31: Diagrama de flujo de la vista Predicción.

V. Discusión y Conclusiones

Esto es un acercamiento para la detección de BPM en la música salsa usando redes neuronales convolucionales. Con la ayuda de las diferentes librerías, los procesos de extracción de datos y análisis de ello son cada vez más fáciles para la comunidad investigativa, ayudando a entender poco a poco como la música y la tecnología pueden estar de la mano.

No solo eso, también podemos ver el alcance que puede aprender las Inteligencias artificiales y de que no solo se reduce a aprender pequeños datos, queda demostrado que no tiene un límite de aprendizaje, lo único que necesita son datos, un buen procesador o GPU y tiempo. En este caso, podríamos decir que aprendió a entender la música salsa, o al menos un abrebocas.

La música es un campo infinito, nunca para de crear nuevos sonidos, nuevos tonos, nuevos ritmos. Cada día se obtienen nuevas canciones, con diferentes patrones y melodías, creciendo de una manera inesperada y sin darnos cuenta todas estas canciones tienen datos interesantes detrás de ellos; deberíamos sorprendernos con la cantidad de variables y campos inexplorados de la música. En esta investigación se tocó un tema en específico de un género en específico; el BPM y la salsa.

Los modelos propuestos en esta investigación no son los únicos, hay infinidad de modelos; se requiere de más personas que continúen con el análisis de estos. Mi objetivo es dar un abrebocas a la gente que quiera apoyar y seguir innovando en este campo.

En un futuro si es posible separar cada instrumento de una canción, es posible que la detección del BPM sea más sencilla; con el cencerro puede ser posible encontrar más fácil el BPM de este género musical salsa, ya que este instrumento a veces está sincronizado literalmente al BPM de la canción.

Agradecimientos

A los profesores Gerardo Sarria, Daniel Gomez y Carlos Arce. Sin su ayuda este proyecto no habría sido posible.

Referencias

[1] Goodwater, “Understanding Spotify: Making Music Through Innovation,” p. www.goodwatercap.com, 2018. [Online]. Disponible en: www.goodwatercap.comhttps://www.goodwatercap.com/thesis/understanding-spotify.

[2] F. Richter, “Chart: Spotify Keeps Apple Music at Arm’s Length | Statista,” 2020. [Online]. Disponible en: https://www.statista.com/chart/8399/spotify-apple-music-paid-subscribers/H. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1985, ch. 4.

[3] R. de Cande, Nuevo Diccionario de la Música, 202.

[4] U. Marchand and G. Peeters, “The Extended Ballroom Dataset,” ISMIR 2016 Late-Breaking Session, pp. 1–3, 2016.

[5] RAE, “Definición hercio.” [Online]. Disponible en: https://dle.rae.es/hercio

[6] L. Roberts, “Understanding the Mel Spectrogram,” marzo, 2020. [Online]. Disponible en: https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53.
[7] H. Schreiber, “Data-Driven Approaches for Tempo and Key Estimation of Music Recordings”. Enero, 2020.
[8] D. Schwerfeger, “How to Easily Process Audio on Your GPU with TensorFlow,” mar 2020. [Online]. Disponible en: https://towardsdatascience.com/how-to-easily-process-audio-on-your-gpu-with-tensorflow-2d9d91360f06.
[9] A. H. Jackson, Machine learning, 1988, vol. 5, no. 2.
[10] D. P. Kingma and J. L. Ba, “Adam: A method for stochastic optimization,” in 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 2015.